A Comparative Eleven-Model Benchmark for Airline Route Profitability Prediction and Seasonal Cost Analysis

Danny Dong and Zhengkang Zhang

***Abstract*—Accurate airline route profitability prediction is critical for fleet allocation, scheduling, pricing, and network planning under volatile operating conditions. This paper presents a comparative machine learning framework for flight-level profit margin prediction using a public airline route profitability dataset containing 7,974 flights across 30 routes in 202[1]4. The study integrates exploratory operational analysis with predictive modeling to investigate the impact of seasonality, route category, demand level, and aircraft assignment on airline profitability. To preserve practical planning utility, direct revenue and total cost variables are intentionally excluded from the final feature set, allowing the models to infer profitability from operational and market-related characteristics rather than accounting identities. Eleven machine learning models are systematically benchmarked, including linear regression, ensemble learning, and gradient-boosting approaches. Experimental results demonstrate that boosted tree models consistently outperform traditional linear methods. Among all compared models, CatBoost achieves the best predictive performance with test R^2 = 0.7042, RMSE = 20.94, and MAE = 15.80, followed closely by Gradient Boosting and HistGradient Boosting. Seasonal analysis further reveals that prediction errors increase from Peak to Low season, indicating that weak and volatile operating environments are more difficult to model accurately. The results confirm that operational features contain sufficient predictive signal for practical profitability forecasting and demonstrate the effectiveness of gradient-boosting methods for airline decision-support applications.**

***Key Words*—airline profitability, route planning, machine learning, CatBoost, gradient boosting, seasonality.**

I. INTRODUCTION

The airline industry operates under high operating costs, volatile fuel prices, fluctuating passenger demand, and increasingly competitive route networks[2]. Airlines must balance revenue generation with operational efficiency while adapting to seasonal demand changes, fleet availability, airport constraints, and network structure. Although some routes appear profitable at an aggregate level, detailed operational analysis often reveals substantial variation caused by factors such as route category, aircraft assignment, demand level, and destination characteristics. Therefore, accurate profitability prediction has become increasingly important for airline route planning, fleet deployment, and strategic decision-making [3].

Traditional airline profitability studies mainly focus on revenue management, demand forecasting, fleet assignment, and network optimization. However, many existing approaches rely on statistical analysis or theoretical optimization models and evaluate only a limited number of predictive algorithms. As a result, the comparative behavior of different machine learning paradigms under the same airline operational environment remains insufficiently explored.

From a machine-learning perspective, airline profitability prediction is a structured tabular learning problem involving nonlinear relationships and complex feature interactions. Ensemble learning methods such as Random Forest [4], Gradient Boosting [5], XGBoost [6], LightGBM [7, 8], and CatBoost [9] have demonstrated strong predictive capability on tabular datasets because of their ability to capture heterogeneous operational patterns [10]. In addition, interpretability methods such as SHAP provide useful tools for understanding feature contributions and improving model transparency [11].

This study develops a flight-level airline profitability prediction framework using operational and network-related variables to predict *Profit_Margin* [12]. Unlike approaches that directly depend on revenue and cost totals, the proposed framework intentionally excludes these accounting variables from the final feature set in order to preserve practical planning utility. Eleven machine learning models are systematically benchmarked, including linear models, bagging ensembles, and gradient-boosting approaches[13]. The framework further combines exploratory data analysis with SHAP-based interpretation to identify the major operational drivers affecting airline profitability.

The main contributions of this paper are summarized as follows:

- A flight-level airline profitability prediction framework is developed using operational and network-related features, enabling data-driven estimation of route profitability from pre-decision planning variables.
- Eleven machine learning models, including linear methods, ensemble models, and gradient boosting approaches, are systematically benchmarked under a unified experimental setting. SHAP-based interpretability analysis is further incorporated to identify key operational drivers of profit margin variation.
- Experimental results demonstrate that boosted tree models consistently outperform traditional linear and bagging-

D. Dong is with the Department of Computer Science, New Jersey City University, New Jersey, 07102, USA, (e-mail: Ddong@njcu.edu).

Z. Zhang is with the Artificial Intelligence and Software College, Liaoning Petrochemical University, Fushun 113001, P. R. China, (e-mail: 2376498865@qq.com).

Corresponding author: Zhengkang Zhang.

based methods, providing accurate and robust predictions for airline profitability across different operational conditions.

The remainder of this paper is organized as follows. Section II reviews related work. Section III introduces the dataset and preprocessing procedures. Section IV presents the methodology and experimental setup. Section V discusses the comparative results and interpretability analysis. Finally, Section VI concludes the paper and outlines future research directions.

II. Problem Description

A. Dataset Description

This study uses a public airline route profitability dataset from Kaggle [14]. The dataset contains 7,974 flight records across 30 routes during the 2024 calendar year. The raw data include operational, revenue, and cost variables such as aircraft type, passenger volume, load factor, flight hours, route category, demand level, season, total revenue, total cost, and *Profit_Margin*.

The prediction target is defined as flight-level *Profit_Margin*, since profit margin better reflects operational efficiency and enables fair comparison across routes of different scales.

B. Data Preprocessing and Feature Engineering

Missing values in *Ancillary_Revenue*, *Catering_Cost*, and *Handling_Cost* are imputed using median values. The *Flight_Date* field is converted into datetime format, from which calendar features such as month and quarter are extracted. Additional operational indicators, including revenue per passenger and cost per passenger, are generated for exploratory analysis.

For model training, the final feature set includes aircraft capacity, passengers, load factor, flight hours, month, quarter, aircraft type, route category, season, demand level, and destination. Direct revenue and cost totals are excluded to prevent target leakage and preserve the practical value of the prediction framework for route planning.

Categorical variables are label encoded. The dataset is divided into training and testing sets using an 80/20 split, and five-fold cross-validation is additionally employed for model evaluation.

C. Exploratory Business Analysis

The dataset reports approximately USD 2.37 billion in total revenue and USD 575 million in total profit, with an average profit margin of 6.19%. Around 66.3% of flights are profitable. Table I summarizes the overall operating statistics.

TABLE I Dataset and business summary of the proposed analysis framework

Metric	Value
Flights analyzed	7,974
Routes analyzed	30
Observation window	2024 full year
Total revenue	USD 2,371,756,240
Total cost	USD 1,796,275,579
Total profit	USD 575,480,661
Average profit margin	6.19%
Profitable flights	66.3%

Substantial variation exists across routes. Long-haul destinations such as FRA, SIN, and CDG achieve average margins above 40%, while several short-haul routes remain strongly unprofitable. Table II summarizes the best and worst-performing routes.

TABLE II Best and worst routes by average profit margin

Destination	Category	Avg. Margin (%)	Profitable Flights (%)
FRA	Long Haul	46.08	100.00
SIN	Long Haul	42.07	99.38
CDG	Long Haul	41.27	100.00
BKK	Long Haul	39.91	99.06
HKG	Long Haul	38.67	99.70
LAX	Long Haul	-15.59	31.55
RUH	Short Haul	-41.20	8.80
JED	Short Haul	-68.19	0.00
AMM	Short Haul	-76.29	0.46
CAI	Short Haul	-83.96	0.00

These route-level differences indicate that profitability is strongly influenced by operational and network structure rather than random variation alone.

D. Seasonal and Route-Level Analysis

Seasonality significantly affects airline performance. Peak season records the highest load factor and average profit margin (17.55%), while Low season shows negative average profitability. Fig. 1 presents the seasonal operating patterns.

Route-category analysis further demonstrates that long-haul routes provide the strongest economic performance, whereas short-haul routes suffer from structurally weak unit economics. Fig. 2 summarizes these differences.

The proposed analysis framework additionally performs route clustering analysis, revealing that profitability differences also exist within the same route category. This result suggests that route-planning decisions should consider both category-level and destination-level characteristics.

E. Prediction Objective

The supervised learning task is formulated as a regression problem:

$$\hat{y}_i = f(x_i), \tag{1}$$

where x_i denotes the operational feature vector and y_i represents the observed *Profit_Margin*. The objective is to estimate profitability outcomes for airline planning and operational decision support.

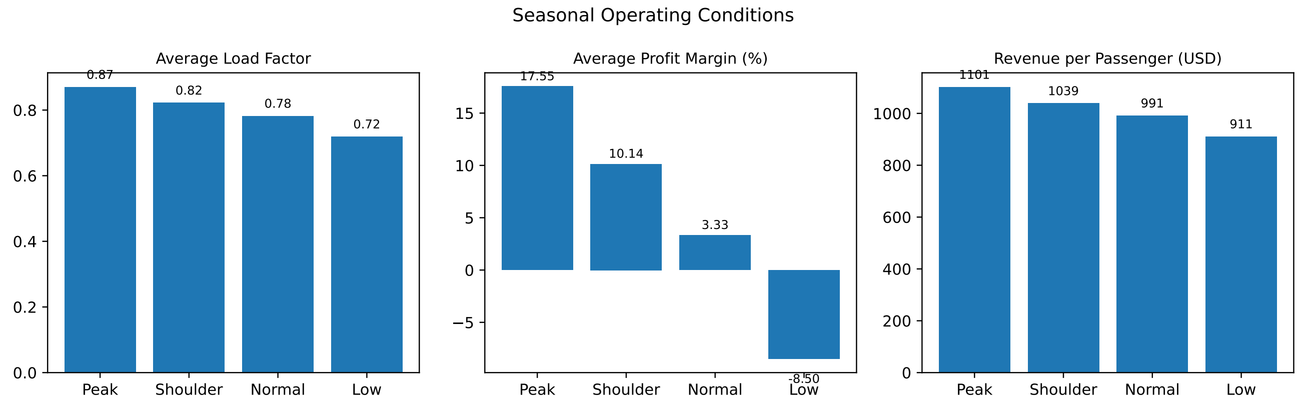

Fig. 1. Seasonal operating conditions. Peak season achieves the highest load factor and profitability, while Low season produces negative average margins.

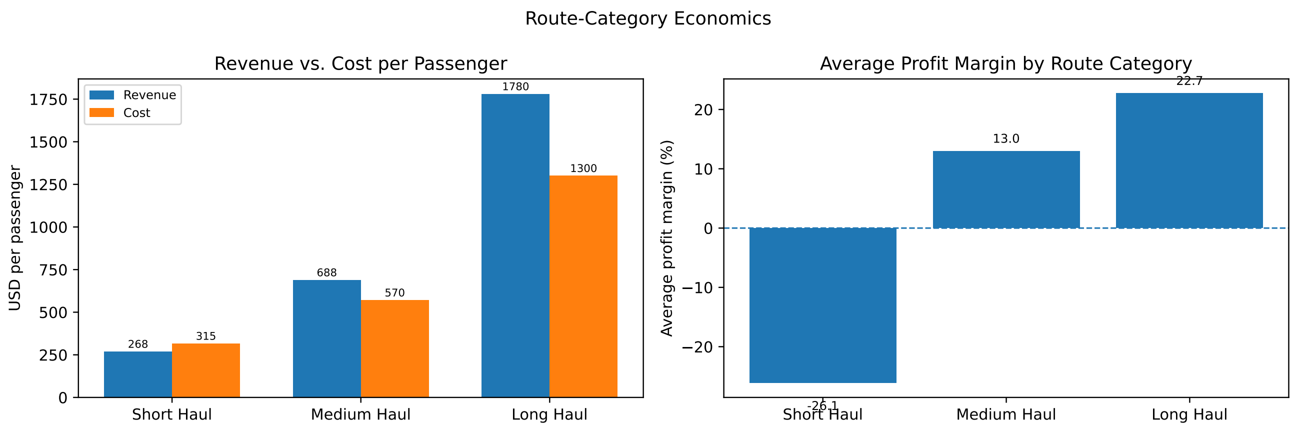

Fig. 2. Route-category economics. Long-haul routes generate the highest profitability, while short-haul routes exhibit negative average margins.

III. Methodology

A. Compared models

The proposed experimental framework compares eleven algorithms spanning several model families. Linear Regression and Ridge Regression serve as transparent baselines and test whether the target is close to a linear combination of the features [15]. Decision Tree provides a single-tree non-linear benchmark [16]. Random Forest and Extra Trees evaluate bagging-based and randomized-tree ensembles [4, 17]. AdaBoost, Gradient Boosting, and HistGradient Boosting represent stage-wise boosting approaches of increasing sophistication [5, 18]. LightGBM and XGBoost serve as high-performance industrial gradient-boosted tree baselines [6, 7]. CatBoost rounds out the comparison with an ordered-boosting implementation that is often competitive on mixed tabular datasets [9].

The important point is that these models are not redundant. A wide benchmark matters because different algorithms encode different assumptions about smoothness, noise, interaction depth, and category handling. If all models were clustered together, the benchmark would not tell much. Instead, the results show meaningful spread, which makes the comparison informative.

B. Evaluation metrics

The experimental framework reports MAE, RMSE, and R^2 on the held-out test set, and one stage also reports cross-validated mean R^2. These metrics capture different aspects of quality. MAE summarizes average absolute miss, RMSE penalizes larger mistakes more heavily, and R^2 measures explained variance relative to the mean predictor. They are defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{4}$$

A strong airline-planning model should perform well across all three metrics, not just one. The model design, train-test evaluation, supervised regression framing, and comparison of predictive models follow standard statistical-learning practice [19, 20, 21, 22].

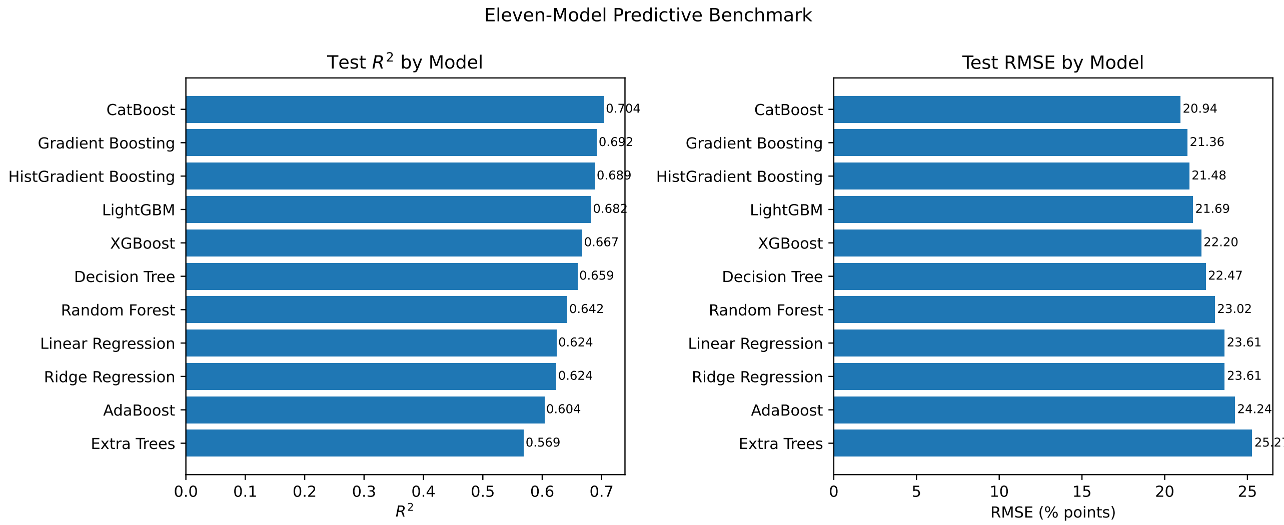

Fig. 3. Eleven-model predictive benchmark on held-out test data. CatBoost achieves the highest R^2 and the lowest RMSE among all compared models.

C. Feature Engineering

The proposed feature design is one of the strongest methodological aspects of the framework. By using calendar, demand, capacity, and route variables while excluding direct accounting totals, it avoids a common mistake in profitability modeling: building a predictor that merely restates known financial outputs. Instead, the model learns to infer expected margin from conditions that are available upstream in planning. This makes the result useful for hypothetical route review. A planner can ask what margin range is plausible for a destination-season-aircraft combination before full realized revenue and cost values are observed.

There are limitations to this setup. Label encoding introduces artificial numeric ordering into categorical variables, and the single 80/20 split may still be optimistic or pessimistic depending on sample composition. Even so, Even so, the proposed framework partially mitigates that issue by reporting five-fold cross-validation mean R^2 in one stage. The results are therefore not just one lucky test split with no stability evidence.

IV. Experimental Results and Discussion

A. Overall Model Comparison

The eleven-model benchmark demonstrates clear performance differences among model families. Fig. 3 and Table III show that CatBoost achieves the best overall performance with test $R^2 = 0.7042$, RMSE = 20.94, and MAE = 15.80. Gradient Boosting and HistGradient Boosting rank closely behind, while LightGBM and XGBoost also achieve competitive results.

Several conclusions can be drawn from the benchmark results. First, linear models already capture a substantial amount of signal, with Linear Regression and Ridge Regression achieving approximately $R^2 = 0.624$. This indicates that operational variables such as load factor, season, and route category are strongly associated with airline profitability.

TABLE III Held-out performance of the eleven models

Model	CV R^2 Mean	Test R^2	Test RMSE	Test MAE
CatBoost	0.7180	0.7042	20.9390	15.7980
Gradient Boosting	0.7079	0.6921	21.3628	15.9886
HistGradient Boosting	0.7087	0.6888	21.4772	16.0302
LightGBM	0.7034	0.6825	21.6926	16.1088
XGBoost	0.6844	0.6675	22.1998	16.3548
Decision Tree	0.6710	0.6593	22.4690	16.6344
Random Forest	0.6758	0.6424	23.0215	16.9378
Linear Regression	0.6456	0.6240	23.6057	18.1587
Ridge Regression	0.6444	0.6237	23.6138	18.1601
AdaBoost	0.6227	0.6037	24.2359	19.9142
Extra Trees	0.6138	0.5691	25.2716	18.4135

Second, boosting methods consistently outperform bagging approaches. Random Forest and Extra Trees improve upon linear baselines but remain weaker than Gradient Boosting, LightGBM, XGBoost, and CatBoost. This suggests that airline profit margin depends heavily on nonlinear interactions and residual correction mechanisms that are better captured by boosting frameworks.

Third, although the performance gap between CatBoost and other boosting models is moderate, CatBoost achieves the best balance across all evaluation metrics. Its stable superiority across cross-validation and held-out testing makes it the strongest candidate for practical deployment.

B. Seasonal Performance Analysis

The proposed framework additionally evaluates seasonal prediction performance. Fig. 4 shows that CatBoost maintains the lowest RMSE across Peak, Shoulder, Normal, and Low seasons.

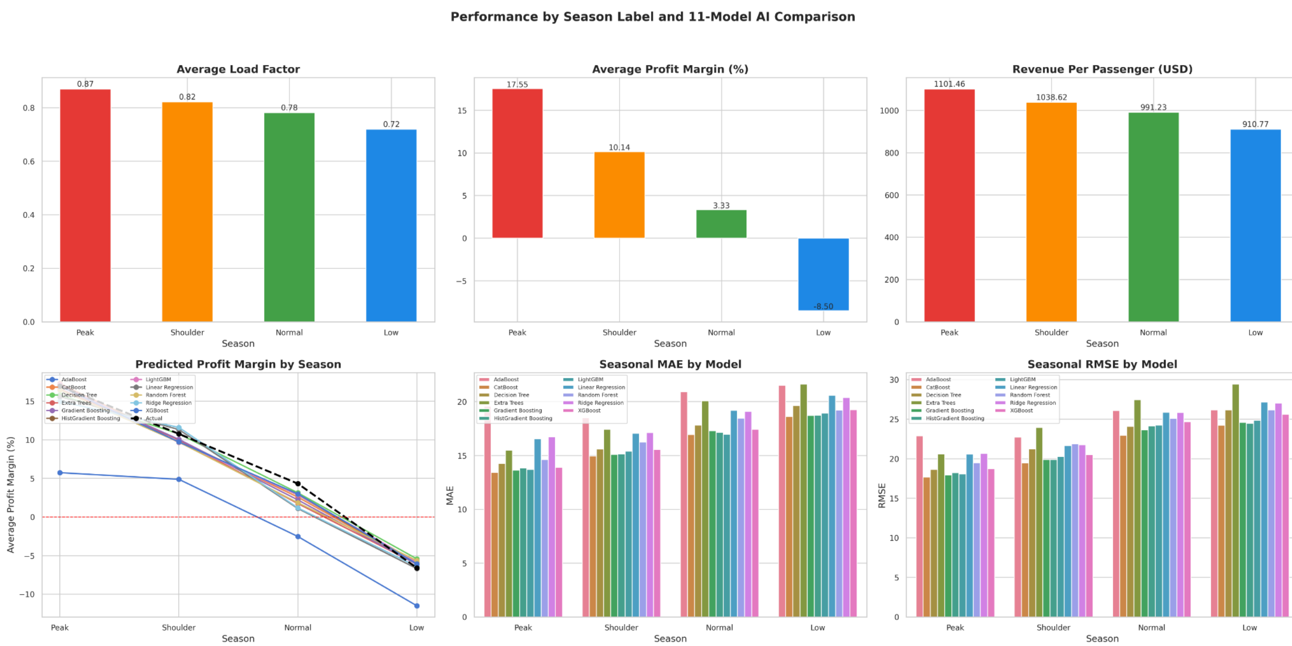

Fig. 5. Seasonal dashboard combining operational KPIs and model performance across different demand periods.

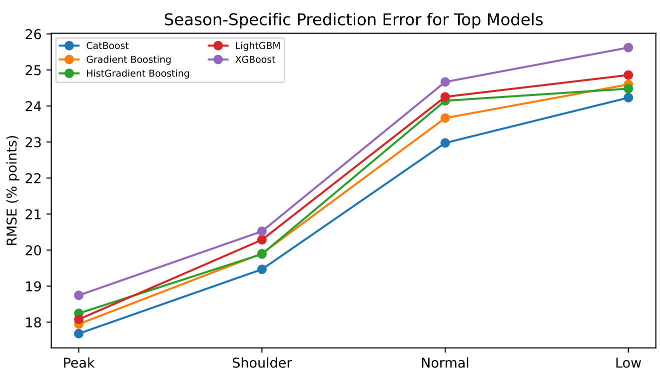

Fig. 4. Season-specific RMSE for the top five models. Prediction errors increase from Peak to Low season.

TABLE IV CatBoost seasonal accuracy versus average seasonal margin

Season	Avg. Margin (%)	MAE	RMSE
Peak	17.55	13.412	17.682
Shoulder	10.14	14.941	19.466
Normal	3.33	16.937	22.974
Low	-8.50	18.624	24.231

The results reveal a clear relationship between business conditions and prediction difficulty. Peak season, characterized by strong demand and positive margins, is easier to predict, whereas Low season produces substantially higher forecasting errors because of weaker and more volatile operating conditions. This seasonal robustness is important for airline deployment because forecasting uncertainty becomes more critical during unfavorable demand periods.

C. Model Interpretability and Deployment Implications

SHAP analysis identifies *Route_Category* as the most influential feature, which is consistent with the exploratory analysis showing strong profitability differences between short-, medium-, and long-haul markets. Other important variables include passenger count, load factor, destination, season, and demand level.

The superior performance of CatBoost can be explained by its ability to model complex interactions among heterogeneous operational variables. Airline profitability is affected by conditional relationships between seasonality, route structure, aircraft assignment, and passenger demand. Boosted-tree models are well suited for capturing these nonlinear effects, whereas linear and bagging approaches are less effective.

Fig. 6 provides an additional comparison of MAE, RMSE, and R^2 across all models.

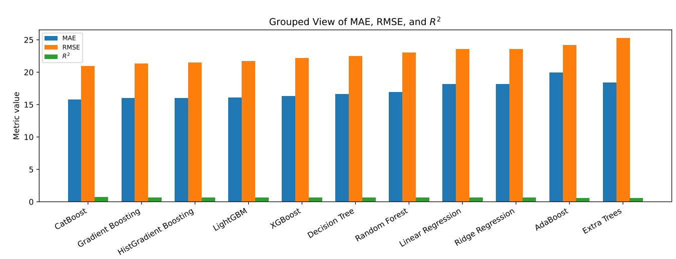

Fig. 6. Grouped comparison of MAE, RMSE, and R^2 across the eleven models.

From an operational perspective, the proposed framework can support multiple airline planning scenarios, including peak-season capacity allocation, low-season risk monitoring, and weak-route restructuring. Since the model relies on operational descriptors rather than realized financial totals, it remains suitable for forward-looking planning and strategic decision support.

V. Conclusion

This study developed a structured machine-learning framework for airline route profitability prediction, demonstrating that route profitability is shaped by strong structural, seasonal, and network-level effects that can be learned effectively from a compact set of operational features. On a dataset of 7,974 flights spanning 30 routes in 2024, the network was profitable overall but exhibited sharp underlying variation: peak season and long-haul routes delivered high margins, while low season and several short-haul markets were systematically weak. The eleven-model benchmark showed that boosted tree methods dominate the predictive task, with CatBoost achieving the best performance (test R^2 = 0.7042, RMSE = 20.94, MAE = 15.80), followed closely by Gradient Boosting, HistGradient Boosting, and LightGBM. A season-specific evaluation further revealed that model errors rise consistently as operating conditions deteriorate from peak to low season, indicating that predictive difficulty is tied to the underlying business environment. These findings establish that the chosen operational feature design, which excludes direct revenue and cost totals, captures enough planning-relevant signal to make profitability prediction both feasible and managerially useful, and that gradient-boosting models provide the most reliable tool for this purpose.

Nevertheless, the analysis has several limitations that suggest clear directions for future work. The study is confined to a single calendar year, potentially masking multi-year cycles, structural breaks, or atypical periods; future work should test temporal generalization by training on one interval and validating on later ones. The current label encoding of categorical variables, while simple, introduces an artificial numeric order that may be imperfect even for tree-based models, and future implementations could systematically compare one-hot encoding, target encoding, and CatBoost's native categorical handling. Although cross-validated mean R^2 is reported in one benchmark stage, the evaluation relies primarily on a single train-test split; a more rigorous design would adopt repeated cross-validation or time-aware validation to better capture month-to-month route dynamics. The intentional exclusion of direct revenue and cost totals is a strength for planning utility but also caps the predictive ceiling, suggesting that a two-model framework—one for planning with operational descriptors and one for ex post diagnostics with richer financial inputs—could be beneficial in practice. In addition, because SHAP interpretability was applied more thoroughly to LightGBM than to the final winning model, future work should extend the same depth of explanation to CatBoost and verify the stability of the most important features. Finally, the rich descriptive outputs—such as CASK, RASK, clustering structure, and seasonal route statistics—point toward more advanced modeling opportunities, including route-segment classification, route-cluster policy learning, and multi-objective optimization that jointly balances profit, utilization, and risk. With modest extensions addressing these points, the current framework can evolve into a solid, production-grade foundation for real-world airline network profitability analytics.

References

[1] Y. Zhai, H. Yang, J. Wang, and Y. An, "Towards robust identification of rail short-wavelength irregularities: A multi-modal fusion of vibration and profile data," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 3, pp. 155–164, September 2025.

[2] J. Wang, M. Zhou, X. Guo, and L. Qi, "Multiperiod asset allocation considering dynamic loss aversion behavior of investors," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 1, pp. 73–81, 2018.

[3] Y. Qing and J. Zhao, "Enhanced gold rush optimizer for feature selection: Application in software fault prediction datasets," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 2, pp. 58–68, June 2025.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3146–3154.

[8] Z. Zhang, S. Dai, C. Li, W. Wang, J. Parron, and E. Herrera, "Human-robot collaborative disassembly profit maximization via improved grey wolf optimizer," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 2, pp. 69–79, June 2025.

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 6638–6648.

[10] C. Xu, H. Wei, X. Guo, S. Liu, L. Qi, and Z. Zhao, "Human-robot collaboration multi-objective disassembly line balancing subject to task failure via multi-objective artificial bee colony algorithm," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 1–6, 2020.

[11] S. M. Lundberg, G. G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[12] J. Gu, Z. Guo, J. Wang, L. Qi, S. Qin, and S. Qin, "Optimization of multi-factory remanufacturing processes with shared transportation resources using the alns algorithm," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 1, pp. 1–13, March 2025.

[13] Z. Zhang, X. Guo, M. Zhou, S. Liu, and L. Qi, "Multi-objective discrete grey wolf optimizer for solving stochastic multi-objective disassembly sequencing and line balancing problem," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 682–687.

[14] W. Faheem, "Airline route profitability and cost analysis," Kaggle dataset, 2024, accessed through the uploaded Google Colab workflow.

[15] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[22] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.